PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-218880

(43)Date of publication of application : 19.08.1997

(51)Int.Cl. G06F 17/30

(21)Application number : 08-048356

(22)Date of filing : 09.02.1996

(71)Applicant : DAINIPPON SCREEN MFG CO LTD

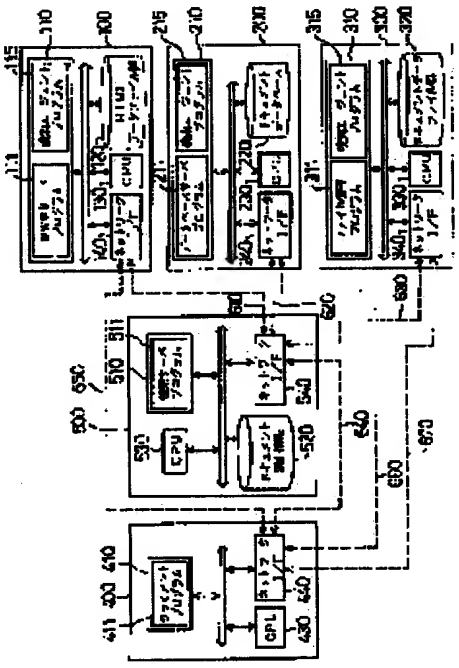
(72)Inventor : KADOMA HISAAKI
KURIHARA DAIKI

(54) DOCUMENT DATA RETRIEVAL SYSTEM

(57)Abstract:

PROBLEM TO BE SOLVED: To turn document data stored somewhere other than a WWW server to a retrieval object as well.

SOLUTION: In the WWW server 100, a data base server 200 and a file server 300, when retrieval agent programs 115, 215 and 315 are respectively activated, CPUs 130, 230 and 330 perform processings corresponding to the programs. The CPUs 130, 230 and 330 extract summary information including bibliography information, a key word and a data storage place, etc., from the stored document data. At the time, the CPUs perform extraction corresponding to a management form for the stored document data and the kind of the document data, etc. The CPUs transfer the extracted summary information from network interfaces 140, 240 and 340 through communication channels 610, 620 and 630 to a retrieval server 500.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1] While storing two or more document data with a client, respectively and offering desired document data according to the demand from said client, respectively Two or more document servers from which the data control gestalt over said document data to store differs mutually, Store the document index information constituted by the summary information of each document data, and said document index information is referred to according to the demand from said client. The retrieval server which searches the storing location of desired document data and offers a retrieval result, It is the document data retrieval system which has even if few, and connects and changes mutually through a communication line. Each document server An extract means to extract the summary information of said document data to these document data to store using the extract technique according to said data control gestalt in the document server concerned, respectively, It is the document data retrieval system equipped with a means for said retrieval server to be based on said transmitted summary information, and to generate or update said document index information, by having a transfer means to transmit said extracted summary information to said retrieval server through said communication line.
[Claim 2] In a document data retrieval system according to claim 1 document at least 1 of said two or more document servers It manages with the 1st data control gestalt accessed per file to said document data to store. Other at least one document server The document data retrieval system characterized by managing with the 2nd data control gestalt which can be accessed about the details of a file to said document data to store.
[Claim 3] It is the document data retrieval system characterized by said 2nd data control gestalt being a data control gestalt by the database management system in a document data retrieval system according to claim 2.
[Claim 4] While storing two or more document data with a client, respectively and offering desired document data according to the demand from said client, respectively Two or more document servers from which the data classification of said document data to store differs mutually, Store the document index information constituted by the summary information of each document data, and said document index information is referred to according to the demand from said client. The retrieval server which searches the storing location of desired document data and offers a retrieval result, It is the document data retrieval system which has even if few, and connects and changes mutually through a communication line. Each document server An extract means to extract the summary information of these document data from said document data using the extract technique according to said data classification of said document data to store, respectively, It is the document data retrieval system equipped with a means for said retrieval server to be based on said transmitted summary information, and to generate or update said document index information, by having a transfer means to transmit said extracted summary information to said retrieval server through said communication line.
[Claim 5] Said document data which document at least 1 of said two or more document servers stores in a document data retrieval system according to claim 4 are the document data retrieval system which is document data of a structured statement document and is characterized by said document data which other at least one document server stores being document data other than said structured statement document.

[Translation done.]

* NOTICES *

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention] This invention relates to the document data retrieval system with which the storing location of desired document data is searched based on document index information by the retrieval server connected to each document server through a communication line, when much document data are distributed and stored in two or more document servers.
[0002]
[Description of the Prior Art] In order to search document data efficiently, it is good to store all document data in one document server, and to manage them intensively. However, since it must stop having to treat a lot of document data when the scale of the organization treating document data becomes large, it cannot manage only by one document server and, so, it is necessary to distribute and store a lot of document data in two or more document servers etc.
[0003] When the user (retrieval person) who needs a certain document data does not know in which document server that document data is stored at this time, a retrieval person has to search whether the target document data are stored for every document server based on bibliographic information, a keyword, etc. using a client. This has a very large burden for a retrieval person.
[0004] Then, in order to solve this problem, the retrieval system using the retrieval server which stored document index information is proposed. As such a retrieval system, there is a retrieval system for the document data offered by WWW (World Wide Web), for example. WWW is structure which offers the document data of a hypertext format. In WWW, the document data of such a hypertext format are stored in a WWW server, and a link can be stretched from one document data to other document data. A retrieval person gets the document data made into the purpose by following the link. However, in WWW, a means to search document data is not offered besides following such a link. Then, in order to compensate this, the retrieval system using the retrieval server mentioned above is proposed variously. for example, "the technical problem in SGML documentation-management-system implementation" (Naoki Inoue: NTT Data Communications Systems Corp.) carried on 24th page - 31st page of "Information Processing Society of Japan, the 2nd time, and the collection of technical communication symposium drafts" ('95.7) -- or "Information Processing Society of Japan -- In the "wide area retrieval system in WWW" (Taketo Tamura, Yoichi Muraoka: Waseda University department of science and engineering) carried by 1-169-170 of the collected works of a national conference" ('95.9), the example of the actual retrieval server for WWW is indicated the 51st time.
[0005] To all the WWW servers on a network, this retrieval server is accessed periodically itself, acquires all the document data (all texts of all pages) stored in each WWW server, respectively, generates document index information based on that acquired data, and stores that document index information. And if a retrieval person accesses the retrieval server using a client, a retrieval server will search the storing location of the document data made into the purpose using the stored document index information, and will tell a retrieval person about the retrieval result.
[0006]
[Problem(s) to be Solved by the Invention] There were the following problems in the retrieval system using the retrieval server in such the former.
[0007] ** A retrieval server accesses only the WWW server on a network, acquires the document data stored in these WWW server, and is generating document index information based on these data. For this reason, in this retrieval system, only the document data stored in the WWW server become a

candidate for retrieval. Therefore, even if a retrieval person is going to search the document data stored in common database servers other than a WWW server etc., they cannot be searched.
[0008] ** In WWW, it is also possible to access the database server of further others through the WWW server using the function of the gateway of the WWW server, after a user accesses a certain WWW server using a client. However, since a retrieval server cannot be accessed to the database server offered by such the gateway, the document data stored in such a database server do not become a candidate for retrieval, either.
[0009] Therefore, the purpose of this invention solves the trouble of the above-mentioned conventional technique, and is to offer the document data retrieval system which can be made applicable to retrieval also about the document data stored in addition to the WWW server.
[0010]
[The means for solving a technical problem, and its operation and effectiveness] In order to attain a part of above-mentioned purpose [at least], the 1st invention While storing two or more document data with a client, respectively and offering desired document data according to the demand from said client, respectively Two or more document servers from which the data control gestalt over said document data to store differs mutually, Store the document index information constituted by the summary information of each document data, and said document index information is referred to according to the demand from said client. The retrieval server which searches the storing location of desired document data and offers a retrieval result, It is the document data retrieval system which has even if few, and connects and changes mutually through a communication line. Each document server An extract means to extract the summary information of said document data to these document data to store using the extract technique according to said data control gestalt in the document server concerned, respectively, Having a transfer means to transmit said extracted summary information to said retrieval server through said communication line, said retrieval server makes it a summary to have a means to be based on said transmitted summary information, and to generate or update said document index information.
[0011] Here, as document data, binary data, such as text data (HTML data etc. are included), image data, and voice data, etc. are mentioned. Moreover, as summary information of document data, bibliographic information, such as a title, an implementer, and a creation date, a keyword, data classification, the storing location of document data, etc. are mentioned.
[0012] Thus, in the 1st invention, the data control gestalt over the document data to store is equipped with two or more mutually different document servers. And each document server extracts summary information from document data with an extract means using the extract technique according to the data control gestalt in the document server, and transmits the summary information to a retrieval server through a communication line with a transfer means, respectively. On the other hand, in a retrieval server, document index information is generated or updated based on the transmitted summary information.
[0013] Therefore, even if a WWW server is document data stored in the document server from which a data control gestalt differs, summary information is extracted and it is transmitted to a retrieval server, and in a retrieval server, it is based on the summary information, and document index information is generated or updated by the extract technique according to the data control gestalt in the document server. Therefore, it can consider as the candidate for retrieval also about the document data stored in addition to the WWW server.
[0014] In the document data retrieval system of the 1st invention document at least 1 of said two or more document servers It manages with the 1st data control gestalt accessed per file to said document data to store. Other at least one document server It is desirable to manage with the 2nd data control gestalt which can be accessed about the details of a file to said document data to store.
[0015] Furthermore, as for said 2nd data control gestalt, it is desirable that it is a data control gestalt by the database management system.
[0016] Thus, between two or more document servers, even if one is a document server which is managing with the data control gestalt accessed per file to document data like a WWW server to store, it can make other one the document server which is managing with the data control gestalt which can be accessed about the details of a file to document data like a database server to store. the case of a database server -- a database management system -- management of document data is performed.
[0017] While the 2nd invention stores two or more document data with a client, respectively and offering desired document data according to the demand from said client, respectively Two or more document servers from which the data classification of said document data to store differs mutually,

Store the document index information constituted by the summary information of each document data, and said document index information is referred to according to the demand from said client. The retrieval server which searches the storing location of desired document data and offers a retrieval result, It is the document data retrieval system which has even if few, and connects and changes mutually through a communication line. Each document server An extract means to extract the summary information of these document data from said document data using the extract technique according to said data classification of said document data to store, respectively, Having a transfer means to transmit said extracted summary information to said retrieval server through said communication line, said retrieval server makes it a summary to have a means to be based on said transmitted summary information, and to generate or update said document index information.

[0018] Thus, the data classification of the document data to store is equipped with two or more mutually different document servers in the 2nd invention. In each document server, using the extract technique according to the data classification of the document data, summary information is extracted from document data and the summary information is transmitted to a retrieval server through a communication line with a transfer means by the extract means, respectively. And in a retrieval server, document index information is generated or updated based on the transmitted summary information.

[0019] Therefore, even if a WWW server is a document server from which the data classification of the document data to store differs, by the extract technique according to the data classification of the document data stored in the document server, summary information is extracted and it transmits to a retrieval server. In a retrieval server, it is based on the summary information, and document index information is generated or updated. Therefore, it can consider as the candidate for retrieval also about the document data stored in addition to the WWW server.

[0020] In the document data retrieval system of the 2nd invention, it is desirable that said document data with which said document data which document at least 1 of said two or more document servers stores are document data of a structured statement document, and store other at least one document server are document data other than said structured statement document.

[0021] Thus, between two or more document servers, even if one is a document server which stores document data of a structured statement document like a WWW server, other one can make it the document server which also stores document data other than a structured statement document like the usual file server.

[0022]

[Embodiment of the Invention] Hereafter, the gestalt of operation of this invention is explained based on an example. Drawing 1 is the explanatory view showing the outline of a document data retrieval system as one example of this invention, and drawing 2 is the block diagram showing the detailed configuration of the document data retrieval system of drawing 1 .

[0023] As shown in drawing 1 or drawing 2 , this document data retrieval system is equipped with the WWW server workstation 100, the database server workstation 200, the file server workstation 300, the client workstation 400, and the retrieval server workstation 500, and they are mutually connected by the communication lines 610-670 on a network.

[0024] The WWW server workstation (it abbreviates to a WWW server hereafter.) 100 is equipped with the program memory 110 which memorizes various programs, the HTML data file group 120 constituted with two or more HTML data, CPU130 which performs various processing actuation according to the program in program memory 110, and the network interface 140 for communicating with other workstations through a network as shown in drawing 2 . Here, HTML data mean the document data written by description language called HTML (Hyper Text Mark-up Language). Each HTML data is stored in storages, such as a hard disk, as a file, respectively, and constitutes the HTML data file group 120. In addition, each HTML data is managed by the file management system in an operating system. Therefore, each HTML data can be accessed only per file.

[0025] In program memory 110, summary information is extracted from the HTML data to store, and the retrieval agent program 115 for transmitting to the retrieval server workstation 500 is remembered to be the WWW server program 111 for referring to the HTML data to store as a program.

[0026] The database server workstation (it abbreviates to a database server hereafter.) 200 is equipped with program memory 210, the document database 220 constituted with two or more document data, CPU230, and the network interface 240 as shown in drawing 2 . Here, the document database 220 is constituted by two or more document data files, and each document data file is further constituted by two or more document data. Moreover, each document data is constituted by others, a title, an

implementor name, etc., respectively. [text] In addition, each document data which constitutes the document database 220 is managed by the database management system (Data Base Management System). Therefore, document data can be accessed about the details of a file.
[0027] The database server program 211 for referring to or updating the document data to store as a program and the retrieval agent program 215 for extracting summary information from the document data to store, and transmitting to the retrieval server workstation 500 are memorized by program memory 210.
[0028] The file server workstation (it abbreviates to a file server hereafter.) 300 is equipped with program memory 310, the document data file group 320 constituted with two or more document data, CPU330, and the network interface 340 as shown in drawing 2 . Here, each document data is stored in storages, such as a hard disk, as a file, respectively, and constitutes the document data file group 320. In addition, since each document data is managed by the file management system in an operating system, each document data can be accessed only per file.
[0029] In program memory 310, summary information is extracted from the document data to store, and the retrieval agent program 315 for transmitting to the retrieval server workstation 500 is remembered to be the file management 311 of OS for referring to the document data to store as a program.
[0030] The client workstation (it abbreviates to a client hereafter.) 400 is equipped with program memory 410, CPU430, and the network interface 440 as shown in drawing 2 . The client program 411 for accessing the WWW server 100, a database server 200, a file server 300, or the retrieval server workstation 500 is memorized as a program by program memory 410.
[0031] Moreover, the retrieval server workstation (it abbreviates to a retrieval server hereafter.) 500 is equipped with program memory 510, the document index information 520, CPU530, and the network interface 540 as shown in drawing 2 .
[0032] While generating or updating the document index information 520 as a program based on the transmitted summary information, the retrieval server program 511 for retrieving the document index information 520 is memorized by program memory 510.
[0033] Now, since the document index information 520 is in the condition of nothing when employing the retrieval server 500 for the first time, in the WWW server 100, a database server 200, and a file server 300, the retrieval agent program 115,215,315 starts and, as for CPU130,230,330, the following processings are performed according to these programs, respectively. That is, summary information including the bibliographic information and the keyword of document data, a data storage location, etc. is extracted from all the stored document data, respectively, and the extracted summary information is transmitted to the retrieval server 500 through a communication line 610,620,630 from a network interface 140,240,340.
[0034] By the way, in the WWW server 100, the database server 200, and the file server 300, as mentioned above, while the management gestalten over the stored document data differ, the classification of the stored document data also differs. Therefore, when CPU of each server extracts summary information according to each retrieval agent program 115,215,315, it is necessary to extract according to a management gestalt, classification of document data, etc. to the stored document data, respectively. Hereafter, the technique of an extract of summary information is explained for every server.
[0035] First, the WWW server 100 is explained. In the WWW server 100, the file of each HTML data is classified and held by the hierarchy by the directory, and is brought together below in a certain directory.

[0036] Drawing 3 is the explanatory view showing an example of the summary information extracted in the WWW server 100 of drawing 1 , and drawing 4 is the explanatory view showing an example of the HTML data which became the radical of the summary information of drawing 3 .
[0037] using the easy program which extracts the character string by which the markup is carried out with the specific tag, since document structure is prescribed by the mark surrounded by <> before and after calling HTML data a tag, for example, a title ("BB report") is described by the condition of <TITLE>BB report </TITLE>, as shown in drawing 4 -- about a "title", it can obtain easily among the summary information shown in drawing 3 .
[0038] Moreover, among the summary information shown in drawing 3 , it can obtain from the time stump of the file which the file management system in an operating system (OS) has managed, and, similarly can obtain ["creation date"] from the owner name of a file about an "implementer."
[0039] Furthermore, about "data classification", there are an approach of acquiring from the extension of

a file, the approach of reading the contents of the file and carrying out an automatic judging, etc. among the summary information shown in drawing 3 . Moreover, about a "keyword", although it can obtain by extracting the character string which can serve as a keyword out of the text, the extract of such a keyword is realizable by using a Japanese morphological analysis system like JUMAN (analysis system by the Kyoto University engineering division manager tail laboratory and the Nara Institute of Science and Technology Matsumoto laboratory).

[0040] Next, a database server 200 is explained. The structure of storing document data is defined by the document database 220, and since each document data which constitutes the document database 220 is managed by the database management system, an easy program can extract summary information using SQL (Structured Query Language; Structured QueryLanguage) etc.

[0041] Drawing 5 is the explanatory view showing an example of the summary information extracted in the database server 200 of drawing 1 , and drawing 6 is the explanatory view showing an example of an SQL program used in case the summary information of drawing 5 is extracted.

[0042] For example, when extracting summary information as shown in drawing 5 , an SQL program required in order to extract a "title", a "creation date", an "implementer", and a "keyword" from document data becomes as [show / in drawing 6]. Therefore, summary information can be extracted about each document data, respectively by making the retrieval agent program 215 equipped with such an SQL program, and performing this program by CPU230 to each document data in which it is stored by the document database 220.

[0043] Next, a file server 300 is explained. If summary information can be extracted from each document data they-stored and each file can be accessed based on these summary information even if it is only stored like the document data stored in the file server 300 as a file only created on the file system and is not applied especially as a database, it will come to function enough as a database.

[0044] There are text data, binary data, etc. which do not have a format of a fixed form besides text data with a format of a fixed form as document data stored in such a file server 300. Then, the technique of extracting summary information from a format of a fixed form is explained first.

[0045] Drawing 7 is the explanatory view showing an example of text data with a format of a fixed form, and drawing 8 is the explanatory view showing an example of the summary information extracted from the text data of drawing 7 .

[0046] Summary information as shown in drawing 8 can be easily extracted by using the program which extracts required information using the word used as a keyword, the line count in a document, etc., for example to text data with a format of a fixed form as shown in drawing 7 .

[0047] Next, the technique of extracting summary information from text data, binary data, etc. without a format of a fixed form is explained. When extracting summary information from such data, the file name managed by the file management system in an operating system is obtained for a "title", a "creation date" and an owner are obtained for the date as an "implementer", respectively, and "data classification" is further obtained from the extension of a file. Moreover, if it restricts to text data, it is also possible to extract a keyword by the approach described in explanation of the WWW server 100.

[0048] Drawing 9 is the explanatory view showing an example of the summary information extracted from binary data. In drawing 9 , since the extract of a keyword was not completed for binary data, the item of a "keyword" serves as a null.

[0049] Now, if the summary information extracted [in / as mentioned above / the WWW server 100, the database server 200, and the file server 300] is transmitted to the retrieval server 500 through a communication line 610,620,630 as mentioned above, in the retrieval server 500, the retrieval server program 511 will start and CPU530 will perform the following processings according to the program. That is, two or more transmitted summary information is received from a network interface 540, and sequential storing is carried out as document index information 520. The document index information 520 is built by general data **-SU, such as a relational database, and, specifically, can perform the exchange with the retrieval server program 511 by SQL etc.

[0050] Drawing 10 is the explanatory view showing an example of the document index information 520 stored in the retrieval server 500 of drawing 1 . The document index information 520 is generated by carrying out sequential are recording of the summary information (drawing 3 , drawing 5 , drawing 8 , drawing 9) transmitted, respectively from the WWW server 100, the database server 200, and the file server 300, as shown in drawing 10 .

[0051] Next, when a retrieval person searches the document data made into the purpose, in a client 400, a client program 411 starts and CPU430 performs the following processings. That is, if a retrieval person

inputs parts, keywords, etc. of bibliographic information, such as a title of document data to search, an implementer, and a creation date, into a client 400 as retrieval conditions, CPU430 will transmit a retrieval demand and retrieval conditions to the retrieval server 500 through a communication line 640 from a network interface 440.

[0052] In the retrieval server 500, if CPU530 receives a retrieval demand from a network interface 540, it will extract the summary information which fulfills retrieval conditions out of the stored document index information 520. And the extract result is transmitted to a client 400. When two or more summary information which fulfills retrieval conditions exists, narrowing down, ranking attachment, etc. which are seen in a common database may be performed.

[0053] In a client 400, if CPU430 receives the retrieval result, it can be shown to a retrieval person and, thereby, in addition to this, a retrieval person can know the storing location of the document data made into the purpose, and required information.

[0054] Next, if you wish acquisition of the document data which the retrieval person searched, CPU430 will start access to the server in which the document data is stored. For example, supposing the server is a database server 200, if there is access from a client 400, with a database server 200, the database server program 211 starts, and CPU230 will read the document data which correspond out of the document database 220 according to the program, and will transmit them to a client 400. Thus, when CPU430 of a client 400 receives the transmitted document data, a retrieval person can get the document data made into the purpose.

[0055] In addition, in a client 400, if a hyperlink which is performed in the combination of a WWW server and its browser is used in case a retrieval result is shown to a retrieval person, a retrieval person can get the document data made into the purpose, without being conscious of where the server in which document data are stored is. Moreover, when the storing location of the target document data is not a WWW server, and making it display on the browser of WWW or starting the program of dedication for a display by changing text data into HTML data in the case of image data etc., it can display easily.

[0056] As explained above, in this example, summary information can be extracted from the document data stored in each document server using the extract technique according to a data control gestalt or data classification when the retrieval agent program according to a management gestalt, classification of document data, etc. to the stored document data is made to have, respectively and each CPU processes in each document server according to each retrieval agent program. And each extracted summary information is transmitted to the retrieval server 500 through a communication line, and generates or updates the document index information 520 in the retrieval server 500 based on these summary information. therefore, since the summary information of the document data boiled and stored in the database servers 200 and file servers 300 other than WWW server 100 is also contained in this document index information 520, it can consider as the candidate for retrieval also about the document data stored in addition to the WWW server, and the candidate for retrieval can be extended as compared with the former.

[0057] Moreover, a retrieval agent program is applicable if the structure which can extract summary information from document data can be created not only to WWW, a common database, and a file system but to any systems. For example, even if a database new type is built by a certain database server, when a protocol with the database server program of the database server creates the same retrieval agent program, it is extensible to it.

[0058] In addition, this invention can be carried out in various modes in the range which is not restricted to the above-mentioned example or the above-mentioned operation gestalt, and does not deviate from the summary.

[0059] In the above-mentioned example, although the document server connected to the network was three, the WWW server 100, a database server 200, and a file server 300, the number of them may be two and it may be four or more. Similarly, a client also only has one client 400 on a network, about these, there are two or more and the retrieval server 500 does not interfere, although the number of retrieval servers was also one.

[0060] Moreover, although the retrieval server 500 has the composition of having become independent of a document server, in the above-mentioned example, the function of a retrieval server is given and you may make it make the document server serve as a retrieval server in one document server.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

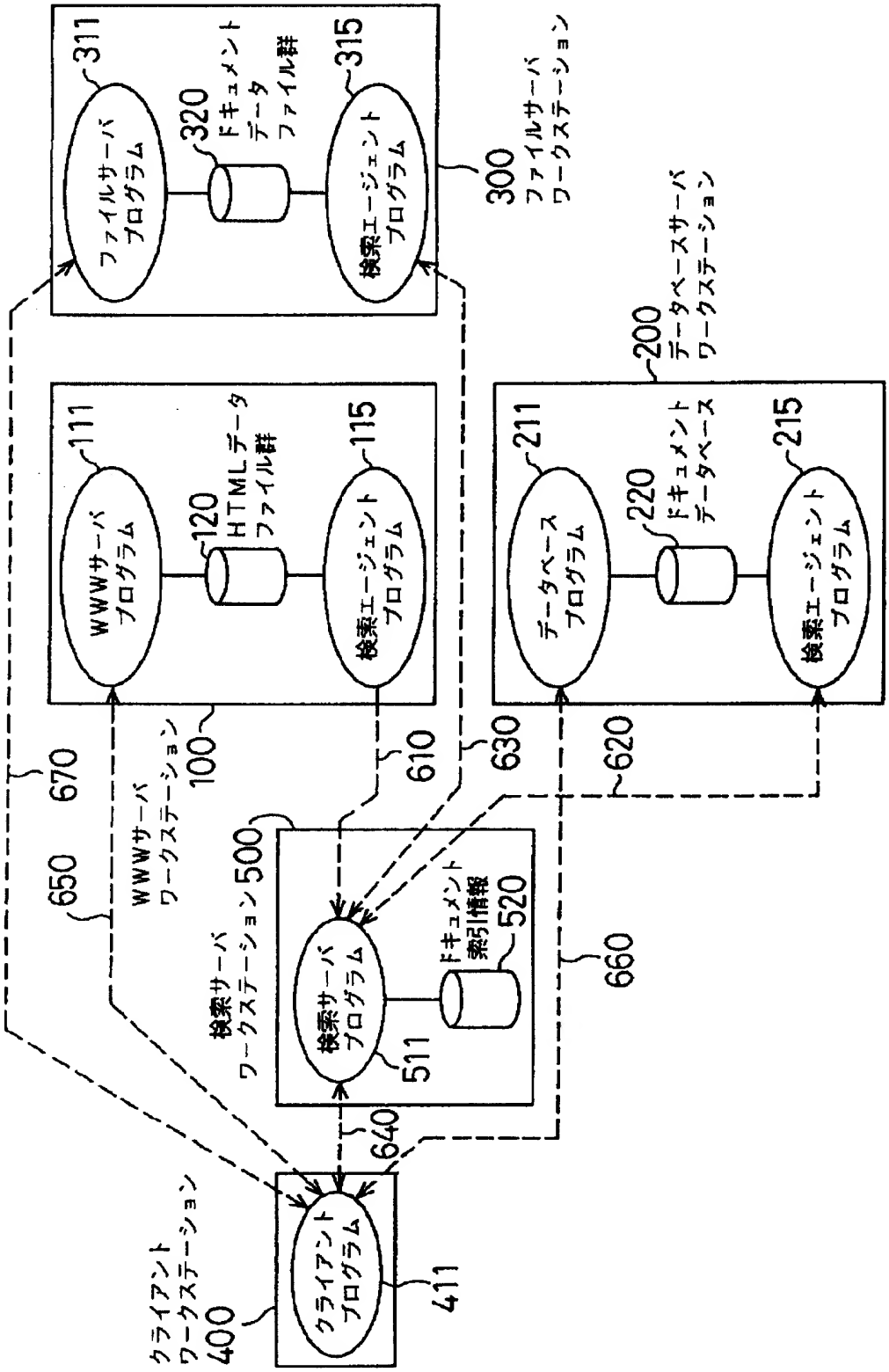
DRAWINGS

[Drawing 3]
タイトル　：ＢＢ報告書
作成年月日　：1995年08月31日
作成者　：大日　太郎
データ種別　：HTML
キーワード　：ネットワーク、分散処理
格納先　：http://www.screen.co.jp/rep.html

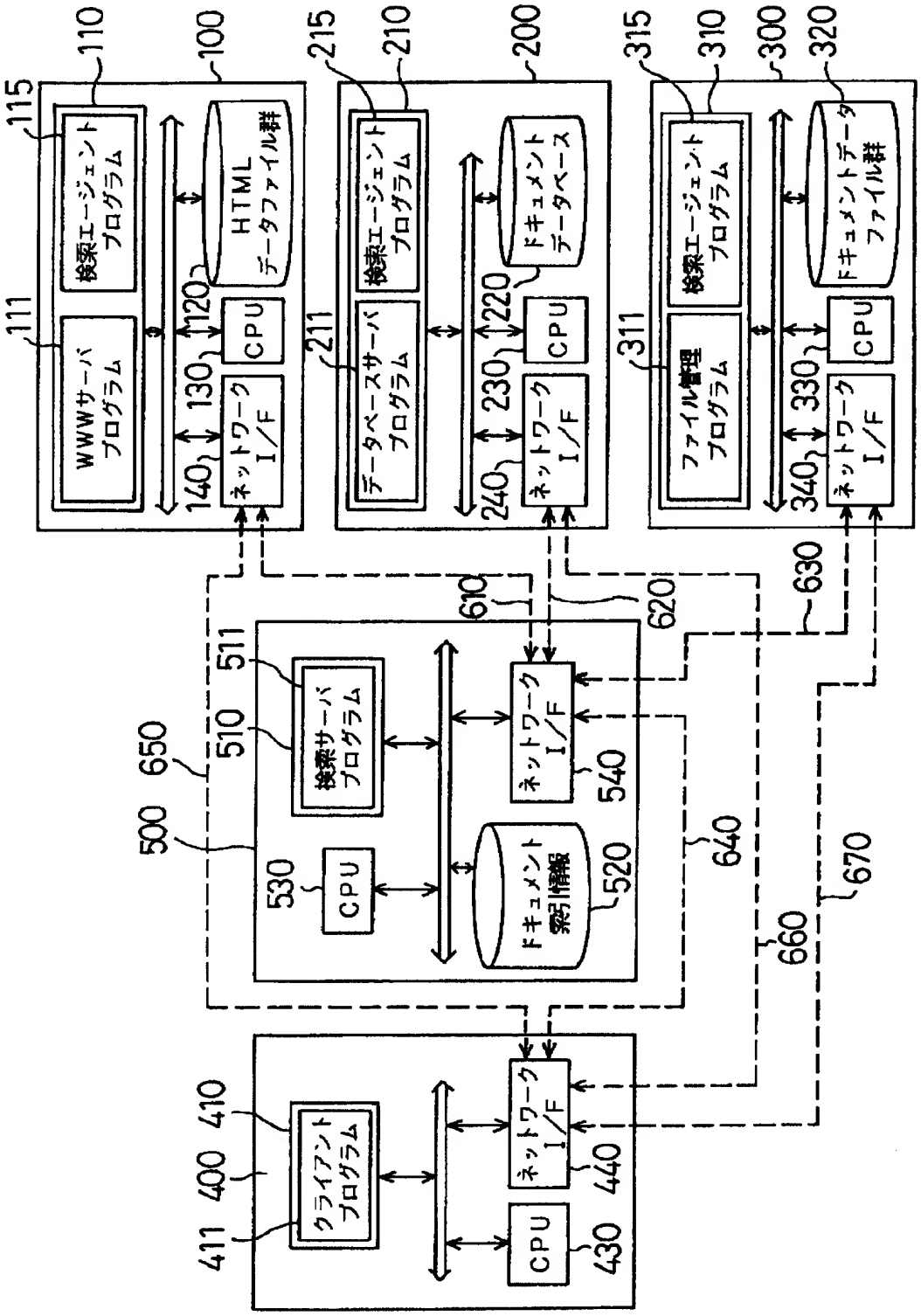
[Drawing 6]
select タイトル,作成年月日,作成者,キーワード
from テーブル名
where ID＝ID値

[Drawing 8]
タイトル　：ＸＸ報告書
作成年月日　：1995年12月5日
作成者　：京都　太郎
データ種別　：テキスト
キーワード　：ＸＸ、ＹＹ
格納先　：ホスト名：ファイル名

[Drawing 1]


311
ファイルサーバプログラム
320
ドキュメントデータファイル群
315
検索エージェントプログラム
300
ファイルサーバワークステーション
200
データベースサーバワークステーション
111
WWWサーバプログラム
120
HTMLデータファイル群
115
検索エージェントプログラム
211
データベースプログラム
220
ドキュメントデータベース
215
検索エージェントプログラム
670
650
WWWサーバワークステーション
100
610
630
620
500
検索サーバワークステーション
検索サーバプログラム
511
ドキュメント索引情報
520
660
640
クライアントワークステーション
400
クライアントプログラム
411

[Drawing 2]


110
100
215
210
200
315
310
300
320
115
検索エージェントプログラム
HTML データファイル群
検索エージェントプログラム
ドキュメントデータベース
検索エージェントプログラム
ドキュメントデータファイル群
111
WWWサーバプログラム
120
CPU
211
データベースサーバプログラム
220
CPU
311
ファイル管理プログラム
CPU
130
ネットワークI/F
230
ネットワークI/F
330
ネットワークI/F
140
240
340
630
610
620
511
510
650
500
検索サーバプログラム
ネットワークI/F
540
530
CPU
ドキュメント索引情報
520
640
670
660
410
400
411
クライアントプログラム
ネットワークI/F
440
CPU
430

[Drawing 4]

```
<HTML>
<HEAD>
<TITLE>ＢＢ報告書</TITLE>
</HEAD>
<BODY>
<H2>ＢＢ報告書</H2>
<HR>
<PRE>
                                        1995年9月5日        大日　太郎

目的：ＸＸのＹＹに関する振る舞いを調査する。
概要及び結論：分散処理に関する実験を行ったので報告する。
ＢＢのネットワークは、ＺＺとＹＹより成り立つが、今回の実験では・・・・
・・・・・
その他：特になし。
</PRE>
<HR>
<A HREF="Welcom.html">ホームページへ</A><P>
</BODY>
</HTML>
```

[Drawing 5]

タイトル　　：ＸＸ報告書
作成年月日　：1995年09月05日
作成者　　：日本　一郎
データ種別　：テキスト
キーワード　：ネットワーク、分散処理
格納先　：ホスト名：データベース名：テーブル名：ＩＤ値

[Drawing 7]

ＸＸ実験報告書

1995年12月5日　　京都　太郎

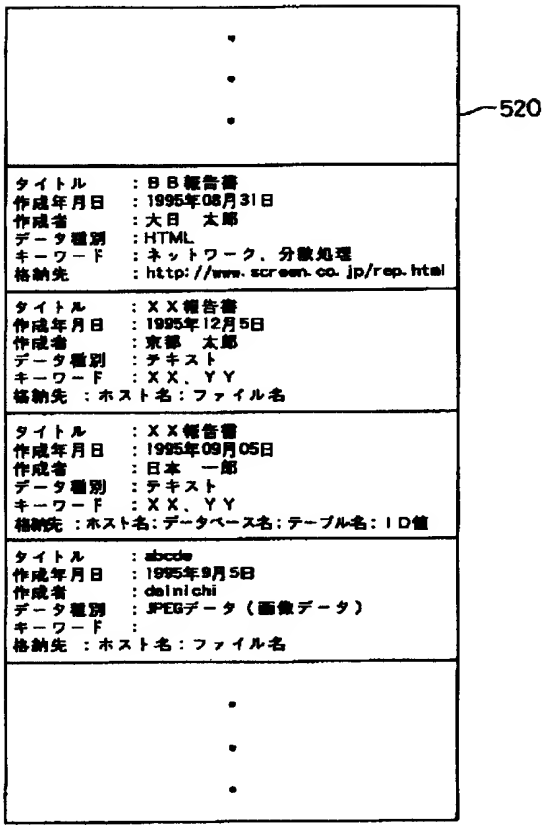
目的：ＸＸのＹＹに関する振る舞いを調査する。

概要及び結論：ＸＸに関する実験を行ったので報告する。
ＸＸは、ＺＺとＹＹより成り立つが、今回の実験では・・・・・
・・・・・
その他：特になし。

[Drawing 9]

タイトル　　：abcde
作成年月日　：1995年9月5日
作成者　　：dainichi
データ種別　：JPEGデータ
キーワード　：
格納先　：ホスト名：ファイル名

[Drawing 10]



[Translation done.]

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号

特開平9－218880

(43)公開日　平成９年(1997)８月19日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　17/30			Ｇ０６Ｆ　15/40	３８０Ｚ	
				３１０Ｃ	
			15/401	３１０Ａ	

審査請求　未請求　請求項の数５　ＦＤ　（全 12 頁）

(21)出願番号　特願平8－48356

(22)出願日　平成８年(1996)２月９日

(71)出願人　000207551
大日本スクリーン製造株式会社
京都府京都市上京区堀川通寺之内上る４丁目天神北町１番地の１

(72)発明者　角間　央章
京都市上京区堀川通寺之内上る４丁目天神北町１番地の１　大日本スクリーン製造株式会社内

(72)発明者　栗原　大樹
京都市上京区堀川通寺之内上る４丁目天神北町１番地の１　大日本スクリーン製造株式会社内

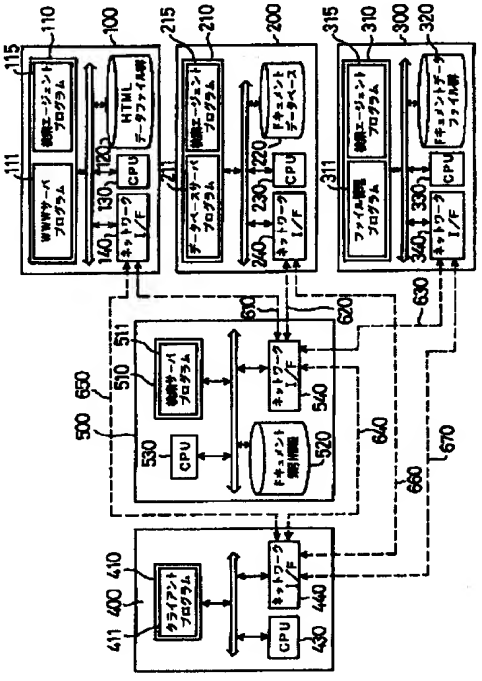
(74)代理人　弁理士　五十嵐　孝雄　（外３名）

(54)【発明の名称】　ドキュメントデータ検索システム

(57)【要約】

【課題】　ＷＷＷサーバ以外に格納されているドキュメントデータについても検索対象とすることができるようにする。

【解決手段】　ＷＷＷサーバ１００、データベースサーバ２００及びファイルサーバ３００では、それぞれ、検索エージェントプログラム１１５，２１５，３１５が起動すると、ＣＰＵ１３０，２３０，３３０はそれらプログラムに従って、処理を行なう。ＣＰＵ１３０，２３０，３３０は格納しているドキュメントデータから書誌情報やキーワードやデータ格納場所などを含むサマリー情報を抽出する。このとき、ＣＰＵは格納しているドキュメントデータに対する管理形態やドキュメントデータの種別などに応じて抽出する。ＣＰＵは抽出したサマリー情報をネットワークインタフェース１４０，２４０，３４０より通信回線６１０，６２０，６３０を介して検索サーバ５００に転送する。



1

【特許請求の範囲】
【請求項１】　クライアントと、
複数のドキュメントデータをそれぞれ格納し、前記クライアントからの要求に応じて所望のドキュメントデータをそれぞれ提供すると共に、格納する前記ドキュメントデータに対するデータ管理形態が互いに異なる２つ以上のドキュメントサーバと、
各ドキュメントデータのサマリー情報によって構成されるドキュメント索引情報を格納し、前記クライアントからの要求に応じて、前記ドキュメント索引情報を参照して、所望のドキュメントデータの格納場所を検索し、検索結果を提供する検索サーバと、
を少なくとも備え、相互に通信回線を介して接続して成るドキュメントデータ検索システムであって、
各ドキュメントサーバは、それぞれ、
当該ドキュメントサーバにおける前記データ管理形態に応じた抽出手法を用いて、格納する前記ドキュメントデータから該ドキュメントデータのサマリー情報を抽出する抽出手段と、
抽出した前記サマリー情報を前記通信回線を介して前記検索サーバに転送する転送手段と、
を備え、
前記検索サーバは、
転送された前記サマリー情報に基づいて前記ドキュメント索引情報を生成または更新する手段を備えるドキュメントデータ検索システム。
【請求項２】　請求項１に記載のドキュメントデータ検索システムにおいて、
前記２つ以上のドキュメントサーバのうちの少なくとも一つのドキュメントサーバは、格納する前記ドキュメントデータに対し、ファイル単位でアクセスする第１のデータ管理形態にて管理を行ない、
他の少なくとも一つのドキュメントサーバは、格納する前記ドキュメントデータに対し、ファイルの細部についてアクセスすることができる第２のデータ管理形態にて管理を行なうことを特徴とするドキュメントデータ検索システム。
【請求項３】　請求項２に記載のドキュメントデータ検索システムにおいて、
前記第２のデータ管理形態は、データベース管理システムによるデータ管理形態であることを特徴とするドキュメントデータ検索システム。
【請求項４】　クライアントと、
複数のドキュメントデータをそれぞれ格納し、前記クライアントからの要求に応じて所望のドキュメントデータをそれぞれ提供すると共に、格納する前記ドキュメントデータのデータ種別が互いに異なる２つ以上のドキュメントサーバと、
各ドキュメントデータのサマリー情報によって構成されるドキュメント索引情報を格納し、前記クライアントからの要求に応じて、前記ドキュメント索引情報を参照して、所望のドキュメントデータの格納場所を検索し、検索結果を提供する検索サーバと、
を少なくとも備え、相互に通信回線を介して接続して成るドキュメントデータ検索システムであって、
各ドキュメントサーバは、それぞれ、
格納する前記ドキュメントデータの前記データ種別に応じた抽出手法を用いて、前記ドキュメントデータから該ドキュメントデータのサマリー情報を抽出する抽出手段と、
抽出した前記サマリー情報を前記通信回線を介して前記検索サーバに転送する転送手段と、
を備え、
前記検索サーバは、
転送された前記サマリー情報に基づいて前記ドキュメント索引情報を生成または更新する手段を備えるドキュメントデータ検索システム。
【請求項５】　請求項４に記載のドキュメントデータ検索システムにおいて、
前記２つ以上のドキュメントサーバのうちの少なくとも一つのドキュメントサーバは、格納する前記ドキュメントデータが構造化文書のドキュメントデータであり、
他の少なくとも一つのドキュメントサーバは、格納する前記ドキュメントデータが前記構造化文書以外のドキュメントデータであることを特徴とするドキュメントデータ検索システム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、多数のドキュメントデータを２つ以上のドキュメントサーバに分散して格納している場合に、各ドキュメントサーバに通信回線を介して接続される検索サーバによって、所望のドキュメントデータの格納場所をドキュメント索引情報に基づいて検索するドキュメントデータ検索システムに関するものである。
【０００２】
【従来の技術】ドキュメントデータを効率よく検索するためには、すべてのドキュメントデータを１つのドキュメントサーバに格納し、集中的に管理するのがよい。しかし、ドキュメントデータを扱う組織の規模が大きくなると、大量のドキュメントデータを扱わなければならなくなるため、１つのドキュメントサーバだけでは管理しきれず、それゆえ、大量のドキュメントデータを複数のドキュメントサーバなどに分散して格納する必要がある。
【０００３】このとき、或るドキュメントデータを必要とするユーザ（検索者）が、そのドキュメントデータがどのドキュメントサーバに格納されているかを知らない場合、検索者はクライアントを使って、書誌情報やキーワードなどに基づき、ドキュメントサーバ毎に、目的の

2

ドキュメントデータが格納されているか否かを検索しなければならない。これは検索者にとって大変負担が大きい。

【０００４】そこで、この問題を解決するために、ドキュメント索引情報を格納した検索サーバを用いた検索システムが提案されている。このような検索システムとしては、例えば、WWW（World Wide Web）により提供されるドキュメントデータを対象とした検索システムがある。WWWは、ハイパーテキスト形式のドキュメントデータを提供する仕組みである。WWWでは、このようなハイパーテキスト形式のドキュメントデータをWWWサーバに格納しており、一つのドキュメントデータから他のドキュメントデータへはリンクを張ることができる。検索者は、そのリンクをたどることによって、目的とするドキュメントデータを得る。しかし、WWWでは、このようなリンクをたどること以外には、ドキュメントデータを検索する手段が提供されていない。そこで、これを補うために、前述した検索サーバを用いた検索システムが種々提案されている。例えば、「情報処理学会、第２回、テクニカルコミュニケーションシンポジウム予稿集」（'９５．７）の第２４頁～第３１頁に掲載されている「ＳＧＭＬ文書管理システム実現における課題」（井上直樹：ＮＴＴデータ通信株式会社）や、或いは「情報処理学会、第５１回、全国大会」（'９５．９）の論文集の１－１６９～１７０に掲載されている「WWWにおける広域検索システム」（田村健人，村岡洋一：早稲田大学理工学部）においては、WWWを対象とした実際の検索サーバの例が開示されている。

【０００５】この検索サーバは、ネットワーク上の全てのWWWサーバに対して、自ら定期的にアクセスして、各WWWサーバに格納されている全ドキュメントデータ（全ページの全テキスト）をそれぞれ取得して、その取得したデータを基にドキュメント索引情報を生成し、そのドキュメント索引情報を格納している。そして、検索者がクライアントを使って、その検索サーバにアクセスすると、検索サーバは格納しているドキュメント索引情報を用いて、目的とするドキュメントデータの格納場所を検索し、その検索結果を検索者に知らせる。

【０００６】

【発明が解決しようとする課題】このような従来における検索サーバを用いた検索システムにおいては、次のような問題があった。

【０００７】①検索サーバは、ネットワーク上のWWWサーバのみにアクセスして、それらWWWサーバに格納されているドキュメントデータを取得し、それらデータを基にドキュメント索引情報を生成している。このため、この検索システムにおいては、WWWサーバに格納されているドキュメントデータだけしか検索対象にならない。従って、WWWサーバ以外の一般的なデータベースサーバなどに格納されているドキュメントデータを、

検索者が検索しようとしても検索することはできない。

【０００８】②WWWにおいては、ユーザがクライアントを使って或るWWWサーバにアクセスした上で、そのWWWサーバのゲートウェイの機能を使って、そのWWWサーバを介してさらに他のデータベースサーバにアクセスすることも可能である。しかし、検索サーバは、このようなゲートウェイにより提供されるデータベースサーバに対してはアクセスすることができないので、そのようなデータベースサーバに格納されているドキュメントデータも検索対象にはならない。

【０００９】従って、本発明の目的は、上記した従来技術の問題点を解決し、WWWサーバ以外に格納されているドキュメントデータについても検索対象とすることができるドキュメントデータ検索システムを提供することにある。

【００１０】

【課題を解決するための手段およびその作用・効果】上記した目的の少なくとも一部を達成するために、第１の発明は、クライアントと、複数のドキュメントデータをそれぞれ格納し、前記クライアントからの要求に応じて所望のドキュメントデータをそれぞれ提供すると共に、格納する前記ドキュメントデータに対するデータ管理形態が互いに異なる２つ以上のドキュメントサーバと、各ドキュメントデータのサマリー情報によって構成されるドキュメント索引情報を格納し、前記クライアントからの要求に応じて、前記ドキュメント索引情報を参照して、所望のドキュメントデータの格納場所を検索し、検索結果を提供する検索サーバと、を少なくとも備え、相互に通信回線を介して接続して成るドキュメントデータ検索システムであって、各ドキュメントサーバは、それぞれ、当該ドキュメントサーバにおける前記データ管理形態に応じた抽出手法を用いて、格納する前記ドキュメントデータから該ドキュメントデータのサマリー情報を抽出する抽出手段と、抽出した前記サマリー情報を前記通信回線を介して前記検索サーバに転送する転送手段と、を備え、前記検索サーバは、転送された前記サマリー情報に基づいて前記ドキュメント索引情報を生成または更新する手段を備えることを要旨とする。

【００１１】ここで、ドキュメントデータとしては、テキストデータ（ＨＴＭＬデータなども含む）や、画像データや音声データなどのバイナリデータなどが挙げられる。また、ドキュメントデータのサマリー情報としては、タイトルや作成者や作成年月日などの書誌情報や、キーワードや、データ種別や、ドキュメントデータの格納場所などが挙げられる。

【００１２】このように、第１の発明では、格納するドキュメントデータに対するデータ管理形態が互いに異なる２つ以上のドキュメントサーバを備えている。しかも、各ドキュメントサーバは、それぞれ、抽出手段によって、そのドキュメントサーバにおけるデータ管理形態

に応じた抽出手法を用いて、ドキュメントデータからサマリー情報を抽出し、転送手段によって、そのサマリー情報を通信回線を介して検索サーバに転送する。一方、検索サーバでは、転送されたサマリー情報に基づいてドキュメント索引情報を生成したり、更新したりする。

【００１３】従って、WWWサーバとはデータ管理形態の異なるドキュメントサーバに格納されたドキュメントデータであっても、そのドキュメントサーバにおけるデータ管理形態に応じた抽出手法によってサマリー情報が抽出されて、検索サーバに転送され、検索サーバにおいて、そのサマリー情報に基づきドキュメント索引情報が生成または更新される。よって、WWWサーバ以外に格納されているドキュメントデータについても検索対象とすることができることになる。

【００１４】第１の発明のドキュメントデータ検索システムにおいて、前記２つ以上のドキュメントサーバのうちの少なくとも一つのドキュメントサーバは、格納する前記ドキュメントデータに対し、ファイル単位でアクセスする第１のデータ管理形態にて管理を行ない、他の少なくとも一つのドキュメントサーバは、格納する前記ドキュメントデータに対し、ファイルの細部についてアクセスすることができる第２のデータ管理形態にて管理を行なうことが好ましい。

【００１５】またさらに、前記第２のデータ管理形態は、データベース管理システムによるデータ管理形態であることが好ましい。

【００１６】このように、２つ以上のドキュメントサーバのうち、一つはWWWサーバのような、格納するドキュメントデータに対しファイル単位でアクセスするデータ管理形態にて管理を行なっているドキュメントサーバであっても、他の一つは例えばデータベースサーバのような、格納するドキュメントデータに対しファイルの細部についてアクセスすることができるデータ管理形態にて管理を行なっているドキュメントサーバとすることができる。データベースサーバの場合は、データベース管理システムよってドキュメントデータの管理が行なわれる。

【００１７】第２の発明は、クライアントと、複数のドキュメントデータをそれぞれ格納し、前記クライアントからの要求に応じて所望のドキュメントデータをそれぞれ提供すると共に、格納する前記ドキュメントデータのデータ種別が互いに異なる２つ以上のドキュメントサーバと、各ドキュメントデータのサマリー情報によって構成されるドキュメント索引情報を格納し、前記クライアントからの要求に応じて、前記ドキュメント索引情報を参照して、所望のドキュメントデータの格納場所を検索し、検索結果を提供する検索サーバと、を少なくとも備え、相互に通信回線を介して接続して成るドキュメントデータ検索システムであって、各ドキュメントサーバは、それぞれ、格納する前記ドキュメントデータの前記データ種別に応じた抽出手法を用いて、前記ドキュメントデータから該ドキュメントデータのサマリー情報を抽出する抽出手段と、抽出した前記サマリー情報を前記通信回線を介して前記検索サーバに転送する転送手段と、を備え、前記検索サーバは、転送された前記サマリー情報に基づいて前記ドキュメント索引情報を生成または更新する手段を備えることを要旨とする。

【００１８】このように、第２の発明では、格納するドキュメントデータのデータ種別が互いに異なる２つ以上のドキュメントサーバを備えている。各ドキュメントサーバでは、それぞれ、抽出手段によって、そのドキュメントデータのデータ種別に応じた抽出手法を用いて、ドキュメントデータからサマリー情報を抽出し、転送手段によって、そのサマリー情報を通信回線を介して検索サーバに転送する。そして、検索サーバでは、転送されたサマリー情報に基づいてドキュメント索引情報を生成したり、更新したりする。

【００１９】従って、WWWサーバとは格納するドキュメントデータのデータ種別の異なるドキュメントサーバであっても、そのドキュメントサーバに格納されたドキュメントデータのデータ種別に応じた抽出手法によってサマリー情報を抽出し、検索サーバに転送する。検索サーバでは、そのサマリー情報に基づいてドキュメント索引情報を生成または更新する。よって、WWWサーバ以外に格納されているドキュメントデータについても検索対象とすることができることになる。

【００２０】第２の発明のドキュメントデータ検索システムにおいて、前記２つ以上のドキュメントサーバのうちの少なくとも一つのドキュメントサーバは、格納する前記ドキュメントデータが構造化文書のドキュメントデータであり、他の少なくとも一つのドキュメントサーバは、格納する前記ドキュメントデータが前記構造化文書以外のドキュメントデータであることが好ましい。

【００２１】このように、２つ以上のドキュメントサーバのうち、一つはWWWサーバのような、構造化文書のドキュメントデータを格納するドキュメントサーバであっても、他の一つは例えば通常のファイルサーバのような、構造化文書以外のドキュメントデータをも格納するドキュメントサーバとすることができる。

【００２２】

【発明の実施の形態】以下、本発明の実施の形態を実施例に基づいて説明する。図１は本発明の一実施例としてドキュメントデータ検索システムの概要を示す説明図であり、図２は図１のドキュメントデータ検索システムの詳細な構成を示すブロック図である。

【００２３】図１または図２に示すように、このドキュメントデータ検索システムは、WWWサーバワークステーション１００と、データベースサーバワークステーション２００と、ファイルサーバワークステーション３００と、クライアントワークステーション４００と、検索

サーバワークステーション５００と、を備えており、それらは互いにネットワーク上の通信回線６１０～６７０によって接続されている。

【００２４】WWWサーバワークステーション（以下、WWWサーバと略す。）１００は、図２に示すように、各種プログラムを記憶するプログラムメモリ１１０と、複数のＨＴＭＬデータによって構成されるＨＴＭＬデータファイル群１２０と、プログラムメモリ１１０内のプログラムに従って種々の処理動作を行なうＣＰＵ１３０と、ネットワークを介して他のワークステーションと通信を行なうためのネットワークインタフェース１４０を備えている。ここで、ＨＴＭＬデータとは、ＨＴＭＬ（Hyper Text Mark-up Language）という記述言語で書かれたドキュメントデータを言う。各ＨＴＭＬデータはハードディスクなどの記憶媒体にそれぞれファイルとして格納されていて、ＨＴＭＬデータファイル群１２０を構成している。なお、各ＨＴＭＬデータは、オペレーティングシステムにおけるファイル管理システムによって管理されている。従って、各ＨＴＭＬデータはファイル単位でのみアクセスすることができる。

【００２５】プログラムメモリ１１０には、プログラムとして、格納するＨＴＭＬデータを参照するためのWWWサーバプログラム１１１と、格納するＨＴＭＬデータからサマリー情報を抽出し、検索サーバワークステーション５００に転送するための検索エージェントプログラム１１５が記憶されている。

【００２６】データベースサーバワークステーション（以下、データベースサーバと略す。）２００は、図２に示すように、プログラムメモリ２１０と、複数のドキュメントデータによって構成されるドキュメントデータベース２２０と、ＣＰＵ２３０と、ネットワークインタフェース２４０を備えている。ここで、ドキュメントデータベース２２０は、例えば、複数のドキュメントデータファイルによって構成され、さらに、各ドキュメントデータファイルは複数のドキュメントデータによって構成されている。また、各ドキュメントデータは、それぞれ、本文の他、タイトルや、作成者名などによって構成されている。なお、ドキュメントデータベース２２０を構成する各ドキュメントデータは、データベース管理システム（Data Base Management System）によって管理されている。従って、ファイルの細部についてドキュメントデータにアクセスすることができる。

【００２７】プログラムメモリ２１０には、プログラムとして、格納するドキュメントデータを参照したり、更新したりするためのデータベースサーバプログラム２１１や、格納するドキュメントデータからサマリー情報を抽出し、検索サーバワークステーション５００に転送するための検索エージェントプログラム２１５が記憶されている。

【００２８】ファイルサーバワークステーション（以下、ファイルサーバと略す。）３００は、図２に示すように、プログラムメモリ３１０と、複数のドキュメントデータによって構成されるドキュメントデータファイル群３２０と、ＣＰＵ３３０と、ネットワークインタフェース３４０を備えている。ここで、各ドキュメントデータはハードディスクなどの記憶媒体にそれぞれファイルとして格納されていて、ドキュメントデータファイル群３２０を構成している。なお、各ドキュメントデータは、オペレーティングシステムにおけるファイル管理システムによって管理されているため、各ドキュメントデータはファイル単位でのみアクセスすることができる。

【００２９】プログラムメモリ３１０には、プログラムとして、格納するドキュメントデータを参照するためのＯＳのファイル管理プログラム３１１と、格納するドキュメントデータからサマリー情報を抽出し、検索サーバワークステーション５００に転送するための検索エージェントプログラム３１５が記憶されている。

【００３０】クライアントワークステーション（以下、クライアントと略す。）４００は、図２に示すように、プログラムメモリ４１０と、ＣＰＵ４３０と、ネットワークインタフェース４４０を備えている。プログラムメモリ４１０には、プログラムとして、WWWサーバ１００やデータベースサーバ２００やファイルサーバ３００或いは検索サーバワークステーション５００にアクセスするためのクライアントプログラム４１１が記憶されている。

【００３１】また、検索サーバワークステーション（以下、検索サーバと略す。）５００は、図２に示すように、プログラムメモリ５１０と、ドキュメント索引情報５２０と、ＣＰＵ５３０と、ネットワークインタフェース５４０を備えている。

【００３２】プログラムメモリ５１０には、プログラムとして、転送されたサマリー情報に基づいてドキュメント索引情報５２０を生成したり、更新したりすると共に、そのドキュメント索引情報５２０を検索したりするための検索サーバプログラム５１１が記憶されている。

【００３３】さて、検索サーバ５００を初めて運用する場合、ドキュメント索引情報５２０は無の状態であるので、WWWサーバ１００、データベースサーバ２００及びファイルサーバ３００では、それぞれ、検索エージェントプログラム１１５，２１５，３１５が起動し、ＣＰＵ１３０，２３０，３３０はそれらプログラムに従って、次のような処理を行なう。即ち、格納している全ドキュメントデータから、それぞれ、ドキュメントデータの書誌情報やキーワードやデータ格納場所などを含むサマリー情報を抽出し、その抽出したサマリー情報をネットワークインタフェース１４０，２４０，３４０より通信回線６１０，６２０，６３０を介して検索サーバ５００に転送する。

【００３４】ところで、WWWサーバ１００、データベ

ースサーバ２００及びファイルサーバ３００では、前述したように、格納しているドキュメントデータに対する管理形態が異なると共に、格納しているドキュメントデータの種別も異なっている。従って、各サーバのＣＰＵが各検索エージェントプログラム１１５，２１５，３１５に従ってサマリー情報を抽出する場合、それぞれ、格納しているドキュメントデータに対する管理形態やドキュメントデータの種別などに応じて抽出する必要がある。以下、各サーバ毎にサマリー情報の抽出の手法について説明する。

【００３５】まず、WWWサーバ１００について説明する。WWWサーバ１００では、各ＨＴＭＬデータのファイルはディレクトリで階層に分類されて収容されており、或るディレクトリ以下に集められている。

【００３６】図３は図１のWWWサーバ１００において抽出されるサマリー情報の一例を示す説明図であり、図４は図３のサマリー情報の基になったＨＴＭＬデータの一例を示す説明図である。

【００３７】図４に示すように、ＨＴＭＬデータは、タグと呼ばれる、前後を<>で囲まれたマークで文書構造が規定されており、例えば、タイトル（「ＢＢ報告書」）は<TITLE>ＢＢ報告書</TITLE>という具合に記述されるので、特定のタグでマークアップされている文字列を抜き出すような簡単なプログラムを用いることよって、図３に示すサマリー情報のうち、「タイトル」については容易に得ることができる。

【００３８】また、図３に示すサマリー情報のうち、「作成年月日」については、オペレーティングシステム（ＯＳ）におけるファイル管理システムが管理しているファイルのタイムスタンプから得ることができ、「作成者」についても同じくファイルの所有者名から得ることができる。

【００３９】さらに、図３に示すサマリー情報のうち、「データ種別」に関しては、ファイルの拡張子から得る方法や、ファイルの内容を読み出して自動判定する方法などがある。また、「キーワード」については、本文中よりキーワードとなり得る文字列を抽出することによって得ることができるが、このようなキーワードの抽出は、例えば、ＪＵＭＡＮ（京都大学工学部長尾研究室、奈良先端科学技術大学院大学松本研究室による解析システム）のような日本語形態素解析システムを利用することによって実現できる。

【００４０】次に、データベースサーバ２００について説明する。ドキュメントデータベース２２０ではドキュメントデータを格納する構造が定義されており、ドキュメントデータベース２２０を構成する各ドキュメントデータはデータベース管理システムによって管理されているので、ＳＱＬ（構造化照会言語；Structured QueryLanguage）等を利用して簡単なプログラムにより、サマリー情報を抽出することができる。

【００４１】図５は図１のデータベースサーバ２００において抽出されるサマリー情報の一例を示す説明図であり、図６は図５のサマリー情報を抽出する際に用いるＳＱＬプログラムの一例を示す説明図である。

【００４２】例えば、図５に示すようなサマリー情報を抽出する場合、「タイトル」，「作成年月日」，「作成者」，「キーワード」をドキュメントデータから抽出するために必要なＳＱＬプログラムは図６に示す如くになる。従って、このようなＳＱＬプログラムを検索エージェントプログラム２１５に備えさせて、ＣＰＵ２３０によって、このプログラムを、ドキュメントデータベース２２０に格納されている各ドキュメントデータに対して実行させることにより、各ドキュメントデータについてそれぞれサマリー情報を抽出することができる。

【００４３】次に、ファイルサーバ３００について説明する。ファイルサーバ３００に格納されているドキュメントデータのように、単にファイルシステム上に作成されたファイルとして格納されているだけであって、特にデータベースとして運用されていなくても、それら格納されている各ドキュメントデータからサマリー情報を抽出し、それらサマリー情報を基に各ファイルにアクセスすることができれば、データベースとして十分機能するようになる。

【００４４】このようなファイルサーバ３００に格納されているドキュメントデータとしては、定型のフォーマットを持ったテキストデータの他、定型のフォーマットを持たないテキストデータやバイナリデータなどがある。そこで、まず、定型のフォーマットからサマリー情報を抽出する手法について説明する。

【００４５】図７は定型のフォーマットを持ったテキストデータの一例を示す説明図であり、図８は図７のテキストデータから抽出されたサマリー情報の一例を示す説明図である。

【００４６】図７に示すような定型のフォーマットを持ったテキストデータに対しては、例えば、キーワードとなる単語や文書中の行数などを利用して必要な情報を抽出するプログラムを用いることによって、容易に、図８に示すようなサマリー情報を抽出することができる。

【００４７】次に、定型のフォーマットを持たないテキストデータやバイナリデータなどからサマリー情報を抽出する手法について説明する。このようなデータからサマリー情報を抽出する場合、例えば、オペレーティングシステムにおけるファイル管理システムによって管理されているファイル名を「タイトル」、日付を「作成年月日」、所有者を「作成者」としてそれぞれ得るようにし、さらに、ファイルの拡張子から「データ種別」を得るようにする。また、テキストデータに限るならば、WWWサーバ１００の説明において記述した方法によってキーワードの抽出を行なうことも可能である。

【００４８】図９はバイナリデータから抽出されたサマ

リー情報の一例を示す説明図である。図９では、バイナリデータのため、キーワードの抽出ができなかったので、「キーワード」の項目は空白となっている。

【００４９】さて、以上のようにしてWWWサーバ１００、データベースサーバ２００及びファイルサーバ３００において抽出されたサマリー情報が、前述したように、通信回線６１０，６２０，６３０を介して検索サーバ５００に転送されると、検索サーバ５００では、検索サーバプログラム５１１が起動し、ＣＰＵ５３０がそのプログラムに従って、次のような処理を行なう。即ち、転送された複数のサマリー情報をネットワークインタフェース５４０より受け取って、ドキュメント索引情報５２０として順次格納する。具体的には、ドキュメント索引情報５２０は、リレーショナルデータベース等の一般的なデータベースで構築され、検索サーバプログラム５１１とのやり取りはＳＱＬ等で行なうことができる。

【００５０】図１０は図１の検索サーバ５００に格納されているドキュメント索引情報５２０の一例を示す説明図である。ドキュメント索引情報５２０は、図１０に示すように、WWWサーバ１００、データベースサーバ２００及びファイルサーバ３００よりそれぞれ転送されてきたサマリー情報（図３、図５、図８、図９）を順次蓄積することによって生成される。

【００５１】次に、検索者が目的とするドキュメントデータを検索する場合は、クライアント４００において、クライアントプログラム４１１が起動して、ＣＰＵ４３０が次のような処理を行なう。即ち、検索者が、検索したいドキュメントデータのタイトル、作成者、作成年月日などの書誌情報の一部やキーワードなどを、検索条件としてクライアント４００に入力すると、ＣＰＵ４３０は、検索要求と検索条件をネットワークインタフェース４４０より通信回線６４０を介して検索サーバ５００に転送する。

【００５２】検索サーバ５００では、ＣＰＵ５３０が、検索要求をネットワークインタフェース５４０より受け取ると、格納しているドキュメント索引情報５２０の中から検索条件を満たすサマリー情報を抽出する。そして、その抽出結果をクライアント４００に送信する。検索条件を満たすサマリー情報が複数存在する場合には、一般のデータベースにおいて見られるような絞り込みやランキング付けなどを行なっても良い。

【００５３】クライアント４００では、ＣＰＵ４３０がその検索結果を受信すると、それを検索者に対して提示し、それにより、検索者は目的とするドキュメントデータの格納場所や、その他必要な情報を知ることができる。

【００５４】次に、検索者が検索したドキュメントデータの取得を希望すると、ＣＰＵ４３０はそのドキュメントデータの格納されているサーバにアクセスを開始する。例えば、そのサーバがデータベースサーバ２００であるとすると、データベースサーバ２００では、クライアント４００からアクセスがあると、データベースサーバプログラム２１１が起動し、ＣＰＵ２３０はそのプログラムに従って、ドキュメントデータベース２２０の中から該当するドキュメントデータを読み出して、クライアント４００に転送する。このようにして転送されたドキュメントデータを、クライアント４００のＣＰＵ４３０が受信することによって、検索者は目的とするドキュメントデータを得ることができる。

【００５５】なお、クライアント４００において、検索結果を検索者に対して提示する際に、WWWサーバとそのブラウザの組合せで行なわれているようなハイパーリンクを使用するようにすれば、検索者はドキュメントデータの格納されているサーバがどこにあるかを意識することなく、目的とするドキュメントデータを得ることができる。また、目的とするドキュメントデータの格納場所がWWWサーバでない場合には、テキストデータをＨＴＭＬデータに変換してWWWのブラウザに表示させたり、画像データ等の場合は表示のための専用のプログラムを起動させたりすることにより、容易に表示を行なうことができる。

【００５６】以上説明したように、本実施例では、各ドキュメントサーバに、それぞれ、格納しているドキュメントデータに対する管理形態やドキュメントデータの種別などに応じた検索エージェントプログラムを備えさせ、各ドキュメントサーバにおいて、各々のＣＰＵが各々の検索エージェントプログラムに従って処理をすることにより、データ管理形態やデータ種別に応じた抽出手法を用いて、格納するドキュメントデータからサマリー情報を抽出することができる。そして、抽出された各サマリー情報は通信回線を介して検索サーバ５００に転送され、検索サーバ５００において、それらサマリー情報に基づいてドキュメント索引情報５２０を生成したり、更新したりする。従って、このドキュメント索引情報５２０には、WWWサーバ１００以外のデータベースサーバ２００やファイルサーバ３００にに格納されているドキュメントデータのサマリー情報も含まれるため、WWWサーバ以外に格納されているドキュメントデータについても検索対象とすることができることになり、従来に比較して検索対象を広げることができる。

【００５７】また、検索エージェントプログラムは、WWW、一般データベース、ファイルシステムのみならず、どのようなシステムに対しても、ドキュメントデータからサマリー情報を抽出することができる仕組みが作成できれば適用することができる。例えば、或るデータベースサーバに、新しいタイプのデータベースが構築されたとしても、そのデータベースサーバのデータベースサーバプログラムとのプロトコルが同じ検索エージェントプログラムを作成することにより、拡張が可能である。

【００５８】なお、本発明は上記した実施例や実施形態に限られるものではなく、その要旨を逸脱しない範囲において種々の態様にて実施することが可能である。

【００５９】上記した実施例においては、ネットワークに接続されているドキュメントサーバは、WWWサーバ１００、データベースサーバ２００及びファイルサーバ３００の３つであったが、２つであっても良いし、４つ以上であっても良い。同じく、クライアントもネットワーク上にクライアント４００が１つあるだけであり、検索サーバも検索サーバ５００が１つだけであったが、これらについても、２つ以上あって差し支えない。

【００６０】また、上記した実施例では、検索サーバ５００はドキュメントサーバと独立した構成となっているが、一つのドキュメントサーバの中に、検索サーバの機能を持たせて、そのドキュメントサーバに検索サーバを兼ねさせるようにしても良い。

【図面の簡単な説明】

【図１】本発明の一実施例としてドキュメントデータ検索システムの概要を示す説明図である。

【図２】図１のドキュメントデータ検索システムの詳細な構成を示すブロック図である。

【図３】図１のWWWサーバ１００において抽出されるサマリー情報の一例を示す説明図である。

【図４】図３のサマリー情報の基になったＨＴＭＬデータの一例を示す説明図である。

【図５】図１のデータベースサーバ２００において抽出されるサマリー情報の一例を示す説明図である。

【図６】図５のサマリー情報を抽出する際に用いるＳＱＬプログラムの一例を示す説明図である。

【図７】定型のフォーマットを持ったテキストデータの一例を示す説明図である。

【図８】図７のテキストデータから抽出されたサマリー情報の一例を示す説明図である。

【図９】バイナリデータから抽出されたサマリー情報の一例を示す説明図である。

【図１０】図１の検索サーバ５００に格納されているド＊キュメント索引情報５２０の一例を示す説明図である。

【符号の説明】

１００…WWWサーバ
１１０…プログラムメモリ
１１１…WWWサーバプログラム
１１５…検索エージェントプログラム
１２０…ＨＴＭＬデータファイル群
１３０…ＣＰＵ
１４０…ネットワークインタフェース
２００…データベースサーバ
２１０…プログラムメモリ
２１１…データベースサーバプログラム
２１５…検索エージェントプログラム
２２０…ドキュメントデータベース
２３０…ＣＰＵ
２４０…ネットワークインタフェース
３００…ファイルサーバ
３１０…プログラムメモリ
３１１…ファイル管理プログラム
３１５…検索エージェントプログラム
３２０…ドキュメントデータファイル群
３３０…ＣＰＵ
３４０…ネットワークインタフェース
４００…クライアント
４１０…プログラムメモリ
４１１…クライアントプログラム
４３０…ＣＰＵ
４４０…ネットワークインタフェース
５００…検索サーバ
５１０…プログラムメモリ
５１１…検索サーバプログラム
５２０…ドキュメント索引情報
５３０…ＣＰＵ
５４０…ネットワークインタフェース
６１０～６７０…通信回線

【図３】

```
タイトル　　：ＢＢ報告書
作成年月日　：1995年08月31日
作成者　　　：大日　太郎
データ種別　：HTML
キーワード　：ネットワーク、分散処理
格納先　　　：http://www.screen.co.jp/rep.html
```

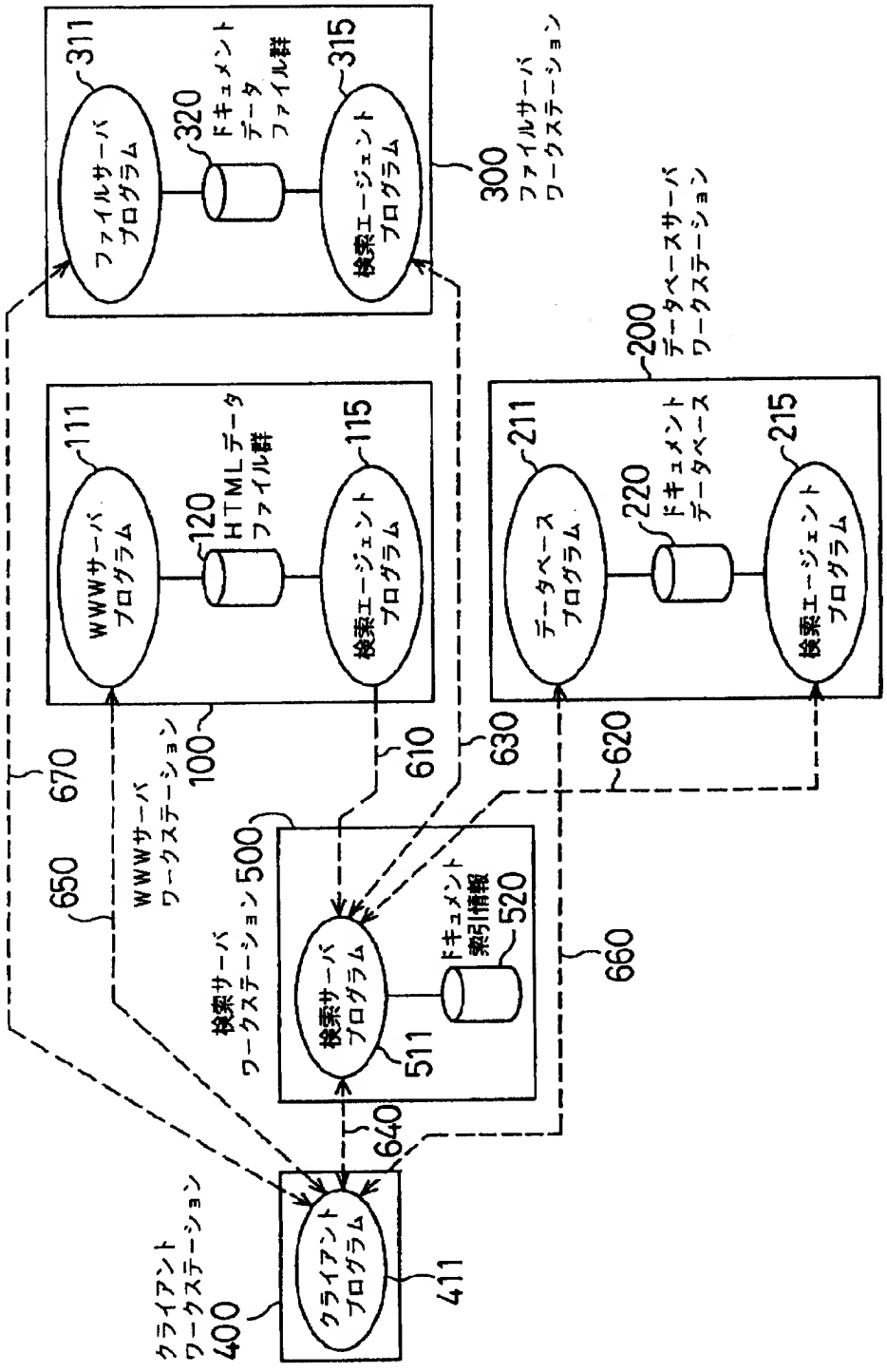
【図６】

```
select タイトル,作成年月日,作成者,キーワード
from テーブル名
where ID＝ID値
```

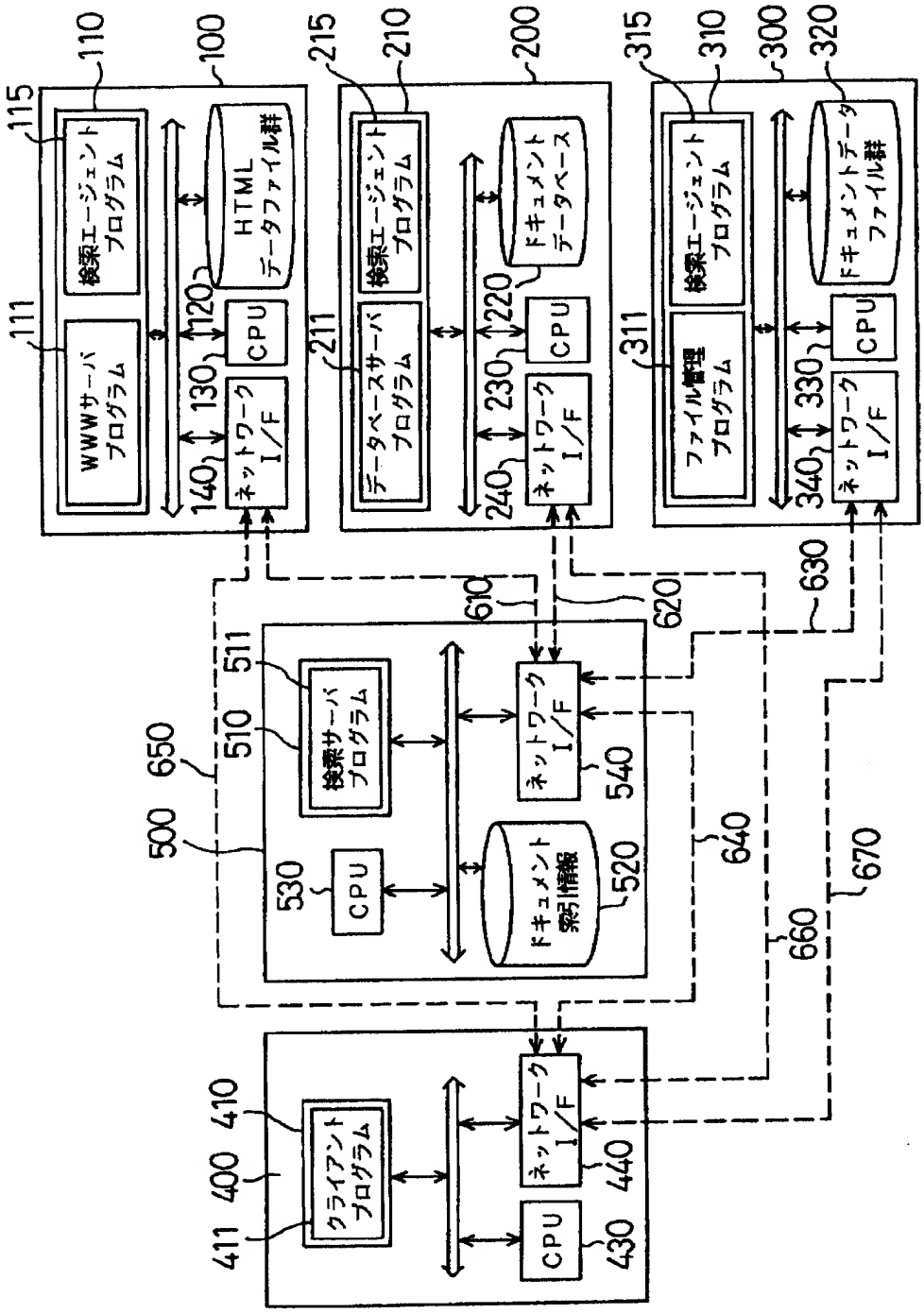
【図８】

```
タイトル　　：ＸＸ報告書
作成年月日　：1995年12月5日
作成者　　　：京都　太郎
データ種別　：テキスト
キーワード　：ＸＸ、ＹＹ
格納先　：ホスト名：ファイル名
```

【図１】


311
ファイルサーバ
プログラム
320
ドキュメント
データ
ファイル群
315
検索エージェント
プログラム
300
ファイルサーバ
ワークステーション
200
データベースサーバ
ワークステーション
111
WWWサーバ
プログラム
120
HTMLデータ
ファイル群
115
検索エージェント
プログラム
211
データベース
プログラム
220
ドキュメント
データベース
215
検索エージェント
プログラム
670
650
WWWサーバ
ワークステーション
100
610
630
620
500
検索サーバ
ワークステーション
検索サーバ
プログラム
511
ドキュメント
索引情報
520
660
640
クライアント
ワークステーション
400
クライアント
プログラム
411

【図２】


WWWサーバプログラム
検索エージェントプログラム
HTMLデータファイル群
CPU
ネットワークI/F
データベースサーバプログラム
ドキュメントデータベース
ファイル管理プログラム
ドキュメントデータファイル群
検索サーバプログラム
ドキュメント索引情報
クライアントプログラム
100
110
111
115
120
130
140
200
210
211
215
220
230
240
300
310
311
315
320
330
340
400
410
411
430
440
500
510
511
520
530
540
610
620
630
640
650
660
670

【図４】

```
<HTML>
<HEAD>
<TITLE>ＢＢ報告書</TITLE>
</HEAD>
<BODY>
<H2>ＢＢ報告書</H2>
<HR>
<PRE>
                                        1995年9月5日　　大日　太郎

目的：ＸＸのＹＹに関する振る舞いを調査する。
概要及び結論：分散処理に関する実験を行ったので報告する。
ＢＢのネットワークは、ＺＺとＹＹより成り立つが、今回の実験では・・・・
・・・・・
その他：特になし。
</PRE>
<HR>
<A HREF="Welcom.html">ホームページへ</A><P>
</BODY>
</HTML>
```

【図５】

タイトル　　：ＸＸ報告書
作成年月日　：1995年09月05日
作成者　　：日本　一郎
データ種別　：テキスト
キーワード　：ネットワーク、分散処理
格納先　：ホスト名：データベース名：テーブル名：ＩＤ値

【図９】

タイトル　　：abcde
作成年月日　：1995年9月5日
作成者　　：dainichi
データ種別　：JPEGデータ
キーワード　：
格納先　：ホスト名：ファイル名

【図７】

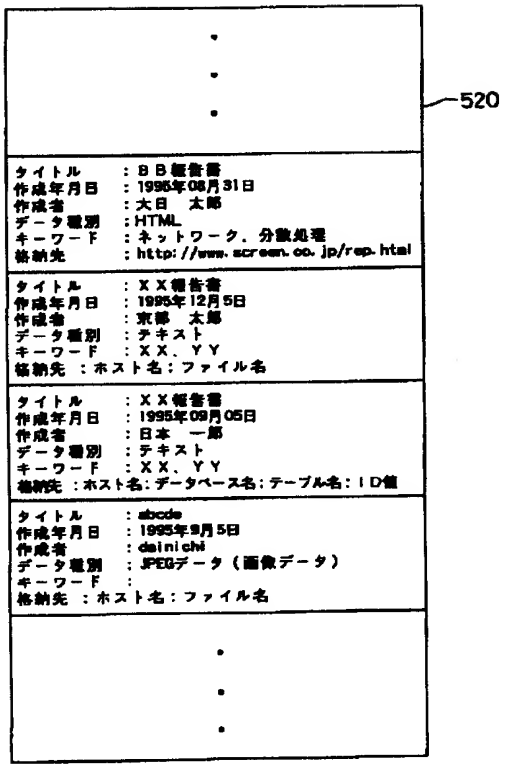
ＸＸ実験報告書

1995年12月5日　　京都　太郎

目的：ＸＸのＹＹに関する振る舞いを調査する。

概要及び結論：ＸＸに関する実験を行ったので報告する。
ＸＸは、ＺＺとＹＹより成り立つが、今回の実験では・・・・・
・・・・・
その他：特になし。

【図１０】


520
タイトル　：ＢＢ報告書
作成年月日　：1995年08月31日
作成者　：大日　太郎
データ種別　：HTML
キーワード　：ネットワーク，分散処理
格納先　：http://www.screen.co.jp/rep.html
タイトル　：ＸＸ報告書
作成年月日　：1995年12月5日
作成者　：京都　太郎
データ種別　：テキスト
キーワード　：ＸＸ、ＹＹ
格納先　：ホスト名：ファイル名
タイトル　：ＸＸ報告書
作成年月日　：1995年09月05日
作成者　：日本　一郎
データ種別　：テキスト
キーワード　：ＸＸ、ＹＹ
格納先　：ホスト名：データベース名：テーブル名：ＩＤ値
タイトル　：abcde
作成年月日　：1995年9月5日
作成者　：dainichi
データ種別　：JPEGデータ（画像データ）
キーワード　：
格納先　：ホスト名：ファイル名